1-9-2015

Interim Report, HD-51897-14, Image Analysis for Archival Discovery (Aida)

Elizabeth M. Lorang
University of Nebraska-Lincoln

Leen-Kiat Soh
University of Nebraska - Lincoln

Interim Report
HD-51897-14
Image Analysis for Archival Discovery (Aida)
Elizabeth Lorang
Leen-Kiat Soh
University of Nebraska-Lincoln
2015-01-09

In the first six months of work on "Image Analysis for Archival Discovery," the project team has made significant strides toward our goal of analyzing more than 7 million newspaper pages in Chronicling America for poetic content. Although we have made some adjustments to our work plan, we remain on task to perform the major research outlined in our proposal.

Activities undertaken from June–November 2014:

- Preparation of training set images (completed)
- Processing of initial data sets to extract/derive features from image data (completed)
- Development of algorithms for describing image characteristics (completed, but future iterations/revisions are likely)
- Training of classifier to recognize poetic content (completed)
- Analyzing of preliminary results and revising of algorithms to achieve higher accuracy rates (completed)
- Processing subset of Chronicling America images (in progress)
- Preparation of program for dividing full page images into image snippets for processing (completed)
- Presentation on preliminary work and results at the Digital Humanities 2014 conference in Lausanne, Switzerland (completed)
- Presentation on project work at Digital Library Federation Forum (completed)
- Recruitment of input from humanities specialists and librarians, archivists, and information professionals outside the advisory board (ongoing)
- Identification of internal and external funding possibilities for future project development (in progress)
- Development of project website, aida.unl.edu (completed)

With the exception of one task—writing the program to interact with the Chronicling America application programming interface—we have accomplished the major milestones outlined in our original work plan for Summer 2014 and a proportional number of milestones outlined for the academic year, relative to the grant period. We have deferred writing the program for interacting with the Chronicling America API until that work becomes necessary for the development of the project. Likely, this work is something we will undertake during the second six months of the grant. Instead, during this period we focused on writing code for creating image snippets from the full newspaper pages, in addition to developing the algorithms for computing visual features and training and testing the classifier.

Over the second six months of the grant period, we will ramp up work to prepare and process image snippets from *Chronicling America*, and these efforts will be the major work of the next six months. In addition, we are currently at work on a journal article setting out the rationale for our project, the algorithms developed, and the accuracy of the classifier through the training

stage. In addition, as outlined in our initial work plan, we will also consult with members of the advisory board and solicit their input for further development.

The only significant deviation from our original plan of work is that we no longer anticipate being able to test our methods on archival materials other than newspapers. According to our original work plan, we intended to develop a partnership with an external institution interested in the application of our methods to archival materials. Moving beyond newspapers in the grant period does not seem likely at this stage, nor is it necessarily the most advantageous research for the present project. We found digital images of newspaper pages to be more varied than anticipated, after reviewing and processing more pages from Chronicling America. Images vary in range effects and contrasts, average intensities for texts and background pixels, resolutions, and to a lesser degree orientations of text columns. These variations have prevented a one-size-fits-all approach to automation and have led to current consideration and experimentation of a hybrid, ensemble approach to automatically identifying poetic content. While it remains viable to transfer the above solution beyond newspapers in the future, it does not seem likely that it is possible during the grant period. Therefore, we are reluctant to develop a relationship with an institution for the purpose of working with their archival materials if we are not able to follow through with this work. In addition, newspapers offer significant additional research areas, so over the next six months we will focus on making connections with scholars interested in the applications of image analysis within newspapers to their own work and to extending the methods beyond *Chronicling America* to other digitized collections of historic newspapers. For example, we have already been in discussion with a scholar interested in songs in historic newspapers (which were printed as poetry), and we are identifying other open collections of historic newspapers that may allow us to extend our work.

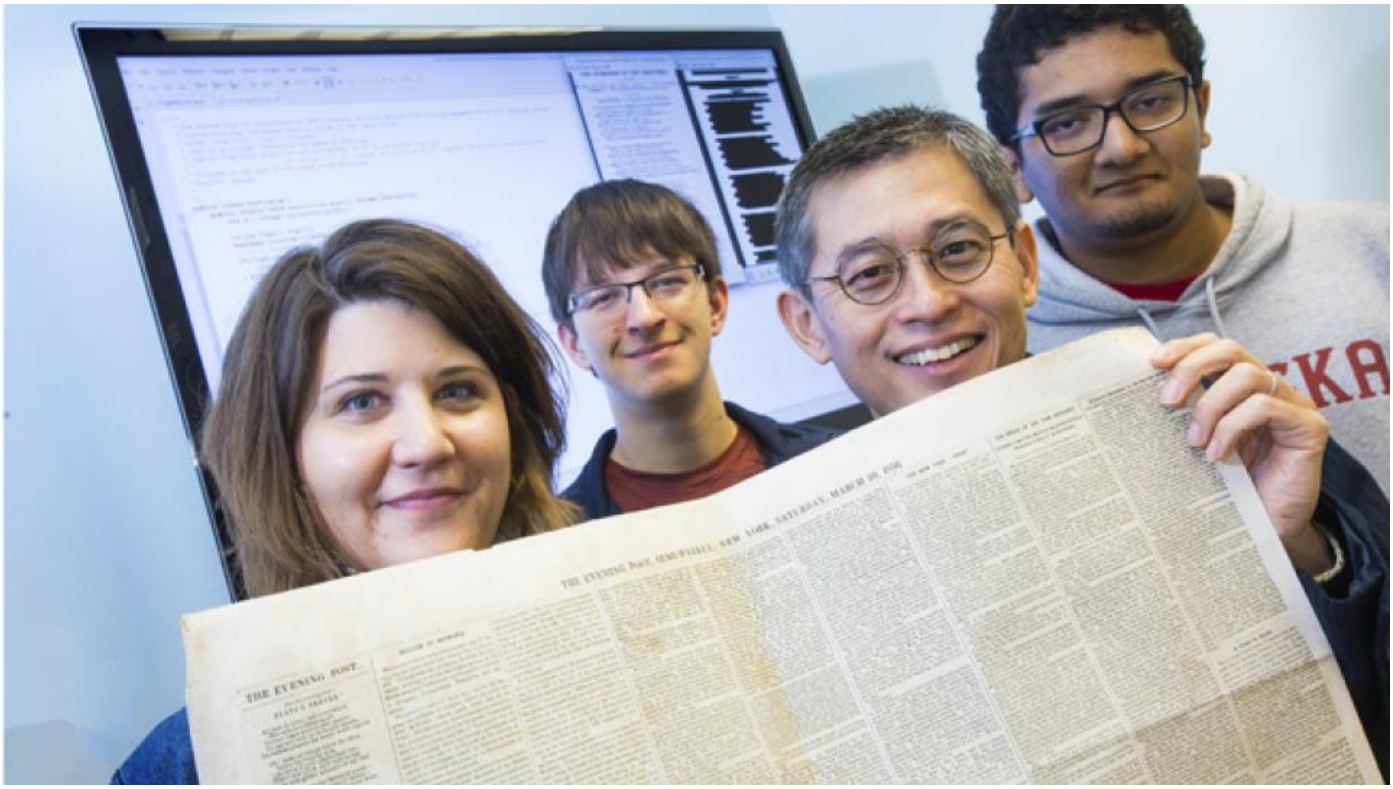
In the first six months of our grant, "Image Analysis for Archival Discovery" was featured in

- An article for UNL Today, the daily online newsletter of the University of Nebraska-Lincoln (http://news.unl.edu/newsrooms/unltoday/article/project-mines-8-million-news-pages-for-poetry/);
- A story aired by the local National Public Radio affiliate, NET, which is available online http://netnebraska.org/article/culture/943643/how-find-poem-200-year-old-newspapers;
- An article feature on Book Patrol (http://bookpatrol.net/mining-newspapers-for-poetry/);
- A blog post on the Poetry Foundation blog, *Harriet* (http://www.poetryfoundation.org/harriet/2014/10/hunting-for-poems-in-a-sea-of-news/).

Appendix: Press Coverage

UNIVERSITY OF NEBRASKA–LINCOLN
UNL Today (http://news.unl.edu/newsrooms/unltoday/)

Project mines 8 million news pages for poetry



Craig Chandler | University Communications

In the upper left corner of a 19th century newspaper is an example of poetry UNL's Elizabeth Lorang (left) is researching. She is collaborating with Leen-Kiat Soh, associate professor of the computer science and engineering, and students Spencer Kulwicki and Maanas Varma Datla, who have developed software to recognize poetry from digitized newspapers.

What differentiates a line of text from a news story and a line of text from a poem? Not much, and that's a problem for researchers of American poetry.

For nearly a century, United States history was documented in newspapers through more than the typical news reports. Millions of poems submitted to and published by newspapers from the late 18th through early 20th centuries also illustrated the lives and concerns of Americans. These poems, if analyzed, could change the history of American literature, said Elizabeth Lorang, research assistant professor and digital humanities projects librarian in the University Libraries.

“Millions of poems were published in newspapers,” she said. “Looking at them will shift the way we understand poetry in the United States.”

But how do researchers locate poems among eight million digitized news pages? Lorang has teamed up with Leen-Kiat Soh, associate professor of computer science and engineering to develop software that will perform image-processing functions to mine data from digital formats.

“This is a big data problem,” Soh said. “What could be done manually, there is now a possibility to do it with computers.”

Similar to text-mining applications, where specific words and phrases are mined from digital sources, the goal of the image processing computer program is to locate specific images or outlines of images. The idea traces back to Lorang’s doctoral dissertation project, when she spent 18 months scouring old newspapers for poems. She was only able to catalog 3,000 poems in that time, but she noticed that the poems were often easily recognizable when looking at the whole page at once.

“Going through the newspapers, I zoomed out on the pages, because the poetic content is so visually distinctive on the page,” Lorang said. “There are differences in white space, justification and other visual cues.”

Lorang and Soh secured a first round of funding to build the software through an 18-month, $60,000 start-up grant from the National Endowment for the Humanities. The funds are enabling Lorang and Soh to work with two undergraduate CSE students to write the code. They plan to have the application running efficiently to locate nearly all of the poems in the millions of digitized news pages of the Chronicling America collection (http://chroniclingamerica.loc.gov).

The students, Spencer Kulwicki and Maanas Varma Datla, joined the project last spring. The UNL sophomores spend about 15 to 20 hours a week working on code and consulting with Lorang and Soh. The experience has been invaluable to them, they said; the interdisciplinary nature of the project is not surprising to the students, but is demonstrative of the possibilities in their chosen career field.

“When I came to UNL, I had no clue what was going on, but at this moment, I can program in three different languages,” Datla said.

Lorang said she believes the project will grow as more researchers learn about it. She also knows that image processing is gaining significant traction in the realm of digital humanities research.

(http://news.unl.edu/sites/default/files/media/poetry-vermont-chron_0.jpg)

Related Links:

Chronicling America (http://chroniclingamerica.loc.

Center for Digital Research in the Humanities (http://cdrh.unl.edu)

Tags:

digital humanities (/free-tags/digital-humanities/)
English (/free-tags/english/)
Computer Science and Engineering (/free-tags/computer-science-and-engineering/)
Center for Digital Research in the Humanities (/free-tags/center-for-digital-research-in-the-humanities/)
College of Arts and Sciences (/free-tags/college-arts-and-sciences/)
Elizabeth Lorang (/free-tags/elizabeth-lorang/)
Leen-Kiat Soh (/free-tags/leen-kiat-soh/)
poetry (/free-tags/poetry/)
newspapers (/free-tags/newspapers/)

“If we think about the massive digital libraries that we’re creating, the tradition has been to use the text that’s created in those processes to enable us to discover content, but at the same time we’re creating digital images,” Lorang said. “If we don’t do anything with those digital images, we’re missing a lot of the potential

of the digital libraries.

“We’re advancing some new methods of digital library research and information retrieval research. Instead of looking for content by trying to discern linguistic features or textual cues, we are looking for the visual forms of those texts, which is a novel approach. I could foresee this software being used in many different applications. We’ve talked about death notices, advertisements, tabular information such as sports scores or weather information. We see this as a methodology that could be relevant to finding a whole variety of content in digitized collections, but we’re starting with the poetry in newspapers as a test case.”

Written by: Deann Gayman | University Communications (/written/deann-gayman-university-communications/)
Published: 10/13/2014

Recent News

2 DAYS AGO

Online programs earn high U.S. News rankings (/newsrooms/unltoday/article/online-programs-earn-high-us-news-rankings/)

2 DAYS AGO

GIS class IDs potential community garden sites (/newsrooms/unltoday/article/gis-class-ids-potential-community-garden-sites/)

2 DAYS AGO

Morrill Hall launches 'Investigate' (/newsrooms/unltoday/article/morrill-hall-launches-investigate/)

How To Find A Poem In 200-year-old Newspapers

by Jackie Sojico, NET Radio

Computer science professor Leen-Kiat Soh goes over the program code with University of Nebraska-Lincoln students Spencer Kulwicki and Manas Varam Datla.

Listen to this story: 00:00 00:00

November 1, 2014 - 8:34am

Liz Lorang is on the hunt for poetry. Not poetry from today, but from 200 years ago. And she's looking for it where it published the most in the 19th century: in American newspapers. She's hoping a team of computer scientists can help her find all of it.

If you've ever looked at microfilm, you know it can be kind of a pain to use even for a few minutes. Imagine going through a century's worth of newspapers…all on microfilm.

"I noticed it in my eyes and my back, and sort of thinking about my posture over time as well. I remember the chairs being so horribly uncomfortable," Lorang said. When Liz Lorang was a graduate student at the University of Nebraska Lincoln, she did exactly that for eight hours a day, for a year and a half. Lorang was doing research for her dissertation about 19th century poetry. She wanted to find as many examples of poetry in newspapers between 1835 and 1880. In a year and a half, she cataloged about 3000. But that's just a tiny fraction of what was published in the entire 19th century.

Poetry as part of daily life

"We tend today to think of it as sort of stodgy old fashioned or esoteric material that only academics engage in. or maybe the kind of you thing you find in hallmark cards at Target," said Amanda Gailey, assistant professor of English at UNL and 19th century American literature expert. Except for American Life in Poetry, a weekly column Ted Kooser started when he was poet laureate, it's hard to find poetry in wide circulation today.

"But in the 19th century when they didn't have movies or television or for the most part even recorded sound, poetry was an important way for them to communicate thoughts, feelings, political opinions, you name it," Gailey said.

Poems in newspapers weren't just written by big names like Walt Whitman and Henry Wadsworth Longfellow. Anyone who was literate and knew how to write poetry could submit a poem to a newspaper.

Lorang keeps a personal collection of old newspapers by her office. "So this is a weekly newspaper from 1841 and the entire first column of the first page is full of poetry. 6 or 8 poems just in this single issue of the newspaper."

- **Want to know where to find poetry in newspapers today? Read Ted Kooser's weekly *American Life in Poetry* column.**
- **Want to see some of the newspapers where Lorang and Soh are looking for poetry? They're from the Library of Congress's**

Teaching a computer to see poetry

Lorang knew there was no way she — or any person — could find all the poems in the archives by themselves.

digital newspaper archives Chronicling America.

“Throughout that time I would go through the microfilm and was thinking that the sorts of cues that I’m looking for as a human reader, the computer should be able to do that same sort of work and be able to do it much more quickly.”

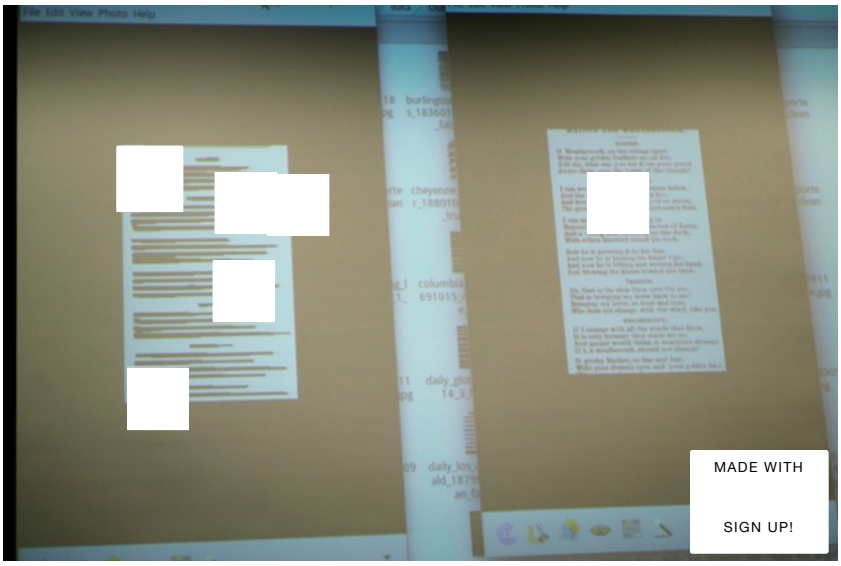
Lorang is now the digital humanities projects librarian at UNL. She turned to Leen-Kiat Soh, a computer science professor also at UNL, with her question. They’re working on building a program now with two UNL students.

Soh and Lorang don't want their program to read the newspapers. Most archives are already doing that using optical character recognition or OCR. OCR is what Google uses to let you search the texts of books online. But text searches only work if you’re looking for a specific word or phrase. Not when you're looking for a kind of text.

Soh and Lorang want their program to recognize the shape of a poem.

What the computer sees

Mouse over the image below to find out. If you're having trouble with the buttons, you can view the image larger. Photo by Jackie Sojico, NET Radio.



“What inspires us is Liz and her students. When they first went through all these newspapers, there’s no way they could read every single page carefully. So they just visually, ah! This looks like a poem, and they zoom in, oh yeah, it’s a poem,” Soh said.

It’s kind of like the way we’re taught to recognize a poem in elementary school. It looks different from other things we read. Soh and Lorang’s program looks for clues like jagged edges instead of straight blocks of text, and more white space, representing line breaks and fewer words, to identify poetry. Right now, the project is still in its early stages.

“Most recently we’re averaging sort of 75 percent. So far the code is really good this thing is not a poem. It’s less good at saying this thing is a poem,” Lorang said.

Soh said that’s because, “human vision is really powerful. We know how to block things out , we how to filter out noises. But to teach computer vision is not easy.”

Eventually they want to the run the program on the 8 million pages archived in the Library of Congress’s digital newspaper collection. Lorang hopes that this project not only makes newspaper archives more accessible to academics, but also changes how we understand poetry’s place in American history — without microfilm.

“The reading that we do in literature courses, we’re exposed to maybe 100 poems that you might read," Lorang said, "we have a very different sense of the history of American poetry than what we get if we actually

Live Jazz at the Vega, Presented by Jazztimesmooth radio

Jan 9 2015 - 7:30pm
Julian Gargiulo

1 of 42 next ›

think about the fact that millions of poems were circulating and people were encountering them and many aspects in their daily life."

Related Articles

- Pain, history & memory: US Poet Laureate muses on creativity
- Nebraska slam poetry competition brings fresh energy to genre
- Nebraska slam poetry competition brings fresh energy to genre
- Teens bare their hearts in poetry
- New Nebraska state poet not afraid to get dirty

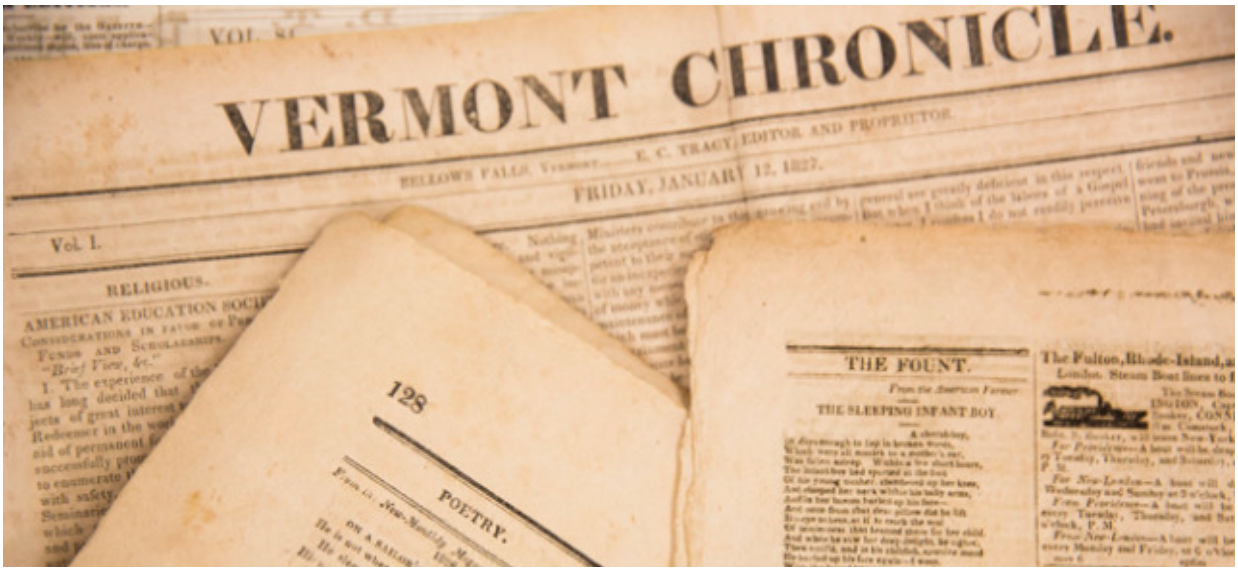
Discussion

Book Patrol: A Haven for Book Culture

Book Patrol is a place where you can share in Michael Lieberman's passion for the printed word, the history of the book as an object and as a cultural artifact.

Mining newspapers for poetry

Posted on October 17, 2014 | **By Michael Lieberman**



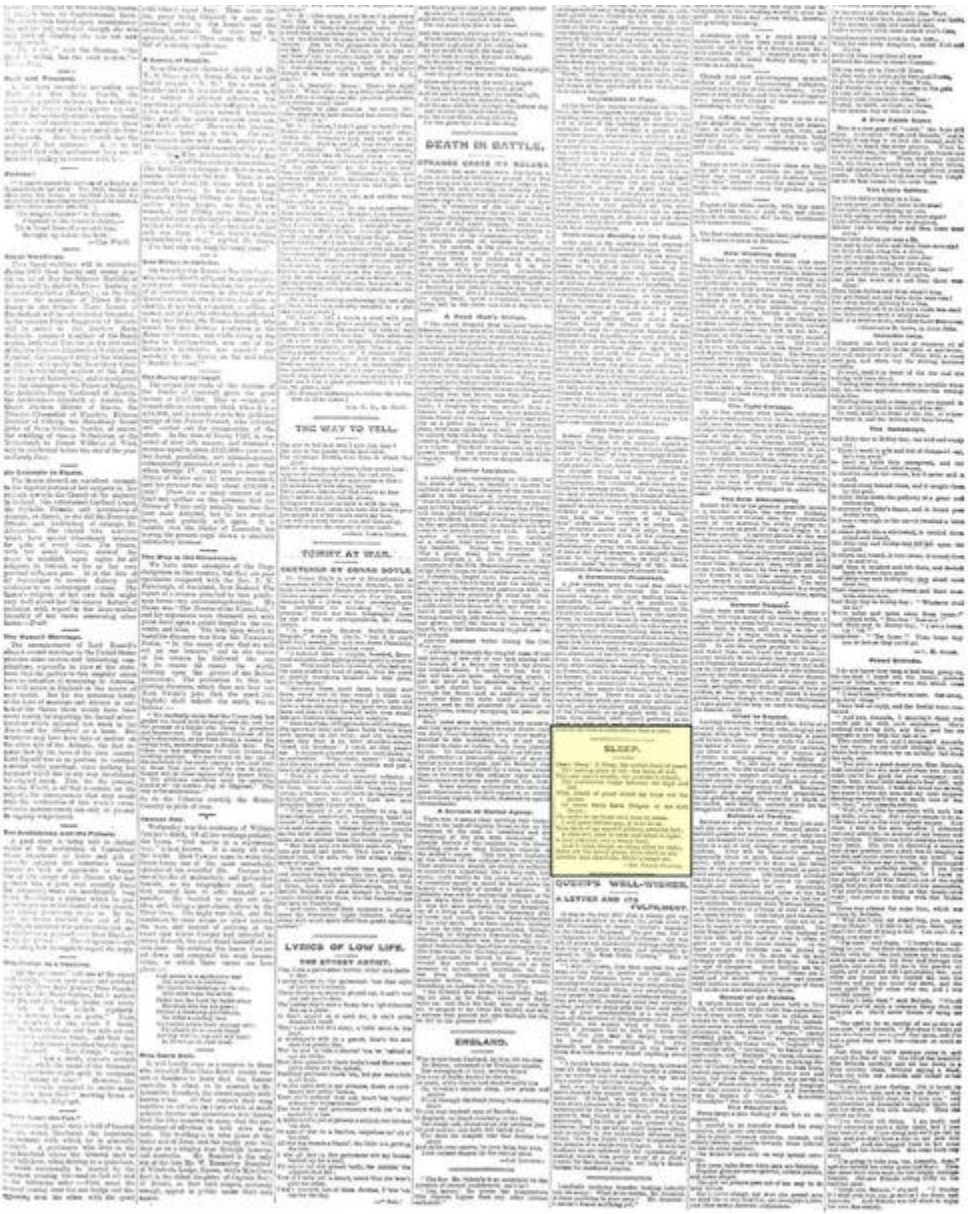
What to do you get when you partner up a digital humanities projects librarian with an associate professor of computer science and engineering?
Answer: Something good.
At the University of Nebraska Elizabeth Lorang, research assistant professor and digital humanities projects librarian in the University Libraries has teamed with Leen-Kiat Soh, associate professor of the computer science and engineering, and a couple of students students to develop software to recognize poetry from digitized newspapers.
“Millions of poems were published in newspapers. Looking at them will shift the way we understand poetry in the United States.” says Lorang.

> Similar to text-mining applications, where specific words and phrases are mined from digital sources, the goal of the image processing computer program is to locate specific images or outlines of images. The idea traces back to Lorang’s doctoral dissertation project, when she spent 18 months scouring old newspapers for poems. She was only able to catalog 3,000 poems in that time, but she noticed that the poems were often easily recognizable when looking at the whole page at once.

In steps Leen-Kiat Soh who views this as” a big data problem” and off they go.
Says Lorang:

> If we think about the massive digital libraries that we’re creating, the tradition has been to use the text that’s created in those processes to enable us to discover content, but at the same time we’re creating digital images. If we don’t do anything with those digital images, we’re missing a lot of the potential of the digital libraries

On the other side of the pond, Andrew Hobbs and Claire Januszewski from the University of Central Lancashire have been keeping a blog focusing on poetry found in nineteenth -century newspapers. The blog, the local press as poetry publisher, 1800-1900, is centered around the hypothesis that “the national network of local newspapers was the largest publisher of nineteenth-century poetry, and the medium through which most encounters with poetry occurred.”



SLEEP.

Come, Sleep! O Sleep, the certain knot of peace,
The baiting place of wit, the balm of woe,
The poor man's wealth, the prisoner's release,
The indifferent judge between the high and low.
With shield of proof shield me from out the prease
Of those fierce darts Despair at me doth throw;
Oh, make in me those civil wars to cease:
I will good tribute pay, if thou do so.
Take thou of me smooth pillows, sweetest bed;
A chamber, deaf to noise and blind to light;
A rosy garland, and a weary head.
And if these things, as being thine by right,
Move not thy heavy grace, thou shalt in me,
Livelier than elsewhere, Stella's image see.

—Sir Philip Sydney.

'Sleep' [Sonnet XXXIX from *Astrophil and Stella*] by Philip Sidney

Hampshire Telegraph April 28 1900

Poetry.

TO DAPHNE.

The stars revolving round heav'n's throne—
  Those eyes of night, that glimmering shine—
Peep timorously and trembling down,
  To catch a brighter glow from thine!

Thy snowy breasts—where nestling Loves
  Delight in soft repose to rest,
Like a young brood of tender doves,
  Hid in their cygnet-downy nest—

May charm with full voluptuous swell;
  But wealth, inestimably worth,
Truth, *Innocence*, beneath them dwell,
  Like gold and diamonds deep in earth!

Sweet bashful morn, with aspect meek,
  And golden locks that glist with dews,
The roseate hue that tints thy cheek
  So delicately, envious views.

Concent'ring beauty's rays in one
  From thee it pours intensely bright—
Thou art the all refulgent sun
  That floods my sky of life with light!

So high exalted—so alone
  In heav'nly charms o'er all thou soars—
Enraptured Nature sees her own
  Divine creation, and adores!

Blackburn, Oct. 13th, 1840. COLIN

'To Daphne', *Blackburn Standard,* 21 Oct 1840
Great stuff.
More on the project:

HARRIET
a poetry blog

CATEGORIES

Featured Blogger
Foundation News
From *Poetry* Magazine
Open Door
Poetry News

FOLLOW HARRIET ON TWITTER

@poetrynews

ABOUT HARRIET

The Poetry Foundation's blog for poetry and related news.

More about Harriet
Contributors
Archive

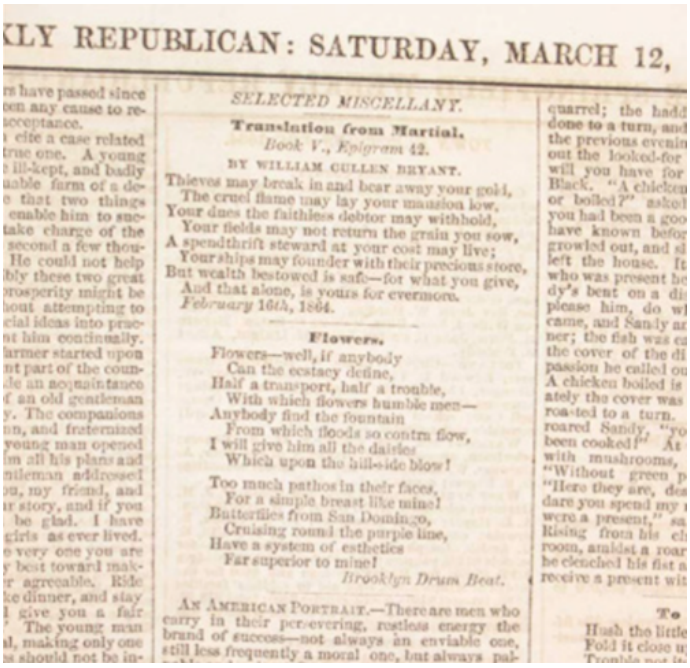
BLOGROLL

Lemon Hound
The Best American Poetry
Bemsha Swing (Jonathan Mayhew)
Dbqp (Geof Huth)
Habenicht Press (David Hadbawnik)
All Links

POETRY NEWS

Hunting for Poems in a Sea of News

BY HARRIET STAFF



KLY REPUBLICAN: SATURDAY, MARCH 12,

SELECTED MISCELLANY.

Translation from Martial.

Book V., Epigram 42.

BY WILLIAM CULLEN BRYANT.

Thieves may break in and bear away your gold,
The cruel flame may lay your mansion low,
Your dues the faithless debtor may withhold,
Your fields may not return the grain you sow,
A spendthrift steward at your cost may live;
Your ships may founder with their precious store,
But wealth bestowed is safe—for what you give,
And that alone, is yours for evermore.

February 16th, 1864.

Flowers.

Flowers—well, if anybody
Can the ecstacy define,
Half a transport, half a trouble,
With which flowers humble men—
Anybody find the fountain
From which floods so contra flow,
I will give him all the daisies
Which upon the hill-side blow!

Too much pathos in their faces,
For a simple breast like mine!
Butterflies from San Domingo,
Cruising round the purple line,
Have a system of esthetics
Far superior to mine!

Brooklyn Drum Beat.

AN AMERICAN PORTRAIT.—There are men who carry in their persevering, restless energy the brand of success—not always an enviable one, still less frequently a moral one, but always pal-

How do you find all the poems fit to print? That's a question Elizabeth Lorang asked in her quest to catalog poetry published in newspapers from the 18th to the 20th century. *UNL Today* reports:

> For nearly a century, United States history was documented in newspapers through more than the typical news reports. Millions of poems submitted to and published by newspapers from the late 18th through early 20th centuries also illustrated the lives and concerns of Americans. These poems, if analyzed, could change the history of American literature, said Elizabeth Lorang, research assistant professor and digital humanities projects librarian in the University Libraries.

Lorang and her colleagues are interested in finding poems that everyday citizens submitted to newspapers, rather than canonical gems that have seen several venues of publication. The question is how to find these poems amid a sea of ink, and for that she's enlisted the help of big data and, of course, some work-study students:

> Lorang has teamed up with Leen-Kiat Soh, associate

> professor of computer science and engineering to develop software that will perform image-processing functions to mine data from digital formats.
>
> "This is a big data problem," Soh said. "What could be done manually, there is now a possibility to do it with computers."
>
> Similar to text-mining applications, where specific words and phrases are mined from digital sources, the goal of the image processing computer program is to locate specific images or outlines of images. The idea traces back to Lorang's doctoral dissertation project, when she spent 18 months scouring old newspapers for poems. She was only able to catalog 3,000 poems in that time, but she noticed that the poems were often easily recognizable when looking at the whole page at once.

Head to *UNL Today* to read more about the project and what the researchers hope to achieve.

Tags: Elizabeth Lorang, UNL Today
Posted in Poetry News on Tuesday, October 14th, 2014 by Harriet Staff.

« Cecily Nicholson's Book-Length Documentary Poem, *From the Poplars*
Impermanent Architecture: At *Jacket 2* Mercedes Eng Explores Built Environments »